Testing Independence When the Form of the Bivariate Distribution is Unspecified

by

Seymour Geisser and Wesley Johnson
University of Minnesota and University of California, Davis

Technical Report 590
October, 1993

*Research supported in part by NIH Grant GM25271

Abstract

We address the problem of testing for independence between X and Y in two situations. In the first we assume that the joint distribution of X and Y is unknown but the observations on X and Y are identifiable. In the second case we assume that the distribution of (X,Y) is exchangeable. Here we consider both when (X,Y) are identifiable and when they are not. For the latter case, an application involving the use of the Hardy-Weinberg law in DNA profiling is given.

Key words: DNA profiling, Exchangeability, Independence, Quantile tables

Testing Independence When the Form of the Bivariate Distribution is Unspecified

Seymour Geisser, University of Minnesota
Wesley Johnson, University of California at Davis

1. Introduction

We provide a methodology for testing independence between X and Y in two situations assuming that the bivariate distribution is continuous. In the first (section 2) we assume that the distribution of (X,Y) is unspecified but the pairs of observations (x_i,y_i) i=1,...,N are identifiable as coming from (X,Y). In the second we assume that (X,Y) is exchangeable. For the latter situation we consider 2 cases: In the first case the pair of observations is identifiable while in the second it is not, as often occurs in same-sex twin data or DNA allelic data from a VNTR locus electrophoretically measured by fragment length size. Here a pair of alleles is obtained from each individual sampled from a well defined population and the paternal and maternal origin of the pair is usually unknown, as distinguished from the first case where the origin is assumed to be known.

The methods devised will be helpful in testing independence at a locus (Hardy-Weinberg equilibrium) and independence between loci (linkage equilibrium).

2. Testing Independence of (X,Y)

Assume a random sample (X_i,Y_i) i=1,...,N is obtained on (X,Y) from a well-defined homogenous population. Calculate the order statistics on X and Y separately obtaining $X_{(1)},X_{(2)},...,X_{(N)}$ and $Y_{(1)},Y_{(2)},...,Y_{(N)}$. Divide the X_i's into q quantile intervals $Q_1,Q_2,...,Q_q$ and similarly for the Y_i's. Form a $q \times q$

quantile table with entries n_{ij}, the number of pairs of (X_i,Y_i) that are in (Q_i,Q_j), i,j=1,...q,

X \ Y	Q_1	Q_2	...	Q_q	
Q_1	n_{11}	n_{12}	...	n_{1q}	N/q
Q_2	n_{21}	n_{22}	...	n_{2q}	N/q
	:	:		:	
Q_q	n_{q1}	n_{q2}	...	n_{qq}	N/q
	N/q	N/q		N/q	N

and for the sake of simplicity assume N is divisible by q with appropriate assignment of pairs where at least one of which is exactly at a quantile boundary. Here for the fixed and equal marginals conditional on the quantiles,

$$E\left(\frac{N_{ij}}{N} \mid H_0\right) = \frac{1}{q^2} \tag{2.1}$$

where H_0 is the hypothesis of indepedence between X and Y. It can be shown that the distribution of

$$X = \sum_{i=1}^{q}\sum_{j=1}^{q} \frac{\left(N_{ij} - \frac{N}{q^2}\right)^2}{\frac{N}{q^2}} = \frac{q^2}{N}\Sigma\Sigma N_{ij}^2 - N \tag{2.2}$$

is asymptotically χ^2 with $(q-1)^2$ degrees of freedom.

For example, when q = 2 so that both the X's and Y's are grouped by their median under independence we randomly pair the X's with the Y's. The number of possible pairings is N!. Now in order to obtain n_{11} of the pairs such that both are less than their respective sample medians we select n_{11} of N/2 X's and n_{11} of the N/2 Y's of those smaller than their respective medians. This is repeated for X's and Y's both above the median. It is also clear from marginal restrictions that $n_{11} = n_{22}$.

Thus the observations below the median and the observations above the median can be matched in $\binom{N/2}{n_{11}}^4 (n_{11}!)^2$ ways. Similarly, $n_{12} = N/2 - n_{11}$ and $n_{21} = N/2 - n_{11}$. These can be matched in $[(N/2 - n_{11})!]^2$ which leads to the hypergeometric probability function

$$\Pr\left[N_{11} = n_{11}\right] = \frac{\binom{N}{2}\binom{\frac{N}{2}}{\frac{N}{2} - n_{11}}}{\binom{N}{\frac{N}{2}}}, \qquad (2.3)$$

$n_{11} = 0,1,\ldots,\frac{N}{2}$.

Further

$$E[N_{11}] = \frac{N}{2} \times \frac{\frac{N}{2}}{N} = \frac{N}{4}$$

$$\mathrm{Var}(N_{11}) = \frac{N^2}{16(N-1)} \doteq \frac{N}{16}$$

and in Appendix I we show that

$$Z = \frac{N_{11} - \frac{N}{4}}{\frac{\sqrt{N}}{4}} \rightarrow N(0,1)$$

and

$$Z^2 \rightarrow \chi_1^2 . \qquad (2.4)$$

The result indicated for general q can be demonstrated in a similar manner. It is clear that the power of such a test depends on the joint distribution of (X,Y) which we leave unspecified. In order to be able to detect dependence if it exists it would be wise then to apply the test for a series of different q in the hope that whatever dependency configuration exists it could be detected for at least one value of q.

3. The Exchangeable Case

Here we assume that the distribution of (X,Y) is exchangeable and one can identify the X's and Y's. The procedure is first to order the set of 2N observations $Z_1, Z_2, \ldots, Z_{2N}$ into q quantile intervals $Q_1, \ldots, Q_q$. Then we form the $q \times q$ quantile table with entries N_{ij}. The only difference being that all 2N observations are used to produce the set of quantiles. It can be shown that under H_0

$$X = \sum_i \sum_j \frac{\left(N_{ij} - \frac{N}{q^2}\right)^2}{\frac{N}{q^2}} = \frac{q^2}{N} \sum_i \sum_j N_{ij}^2 - N \qquad (3.1)$$

is asymptotically χ^2 with q(q–1) degrees of freedom. Here

$$E\left(\frac{N_{ij}}{N} \mid H_0\right) = \frac{1}{q^2} + 0\left(\frac{1}{N}\right) \doteq \frac{1}{q^2}.$$

The result (3.1) for q = 2 is shown in Appendix II.

When (X,Y) is not identifiable, as in same-sex twin data for example, the quantile table is collapsed so that we have entries only along and above the main diagonal,

	Q_1	Q_2		Q_q
Q_1	n_{11}	n_{12}^*	...	n_{1q}^*
		n_{22}	...	n_{2q}^*
			n_{qq}	
Q_q				n_{qq}

where now $n_{ij}^* = n_{ij} + n_{ji}$, since we can only identify the sum of the symmetric entries and not each component. Here one can show that under H_0,

$$X = \frac{q^2}{N} \sum_{i=1}^{q} \left(N_{ii} - \frac{N}{q^2}\right)^2 + \frac{q^2}{2N} \sum\sum_{i<j} \left(N_{ij}^* - \frac{2N}{q^2}\right)^2 \qquad (3.2)$$

is asymptotically χ^2 with $q(q-1)/2$ degrees of freedom, Geisser and Johnson (1992). For the case $q = 2$, the result is shown in Appendix II to be a special case of the identifiable exchangeable situation. For a complete proof, see also Geisser and Johnson (1992).

3. Application

In human DNA forensic identification, several loci obtained from blood or other tissue are probed using electrophoresis to obtain fragment length sizes. These loci have a very large number of presumably discrete alleles. However, the procedure used can neither precisely resolve an allele nor determine the exact number of alleles at a locus. Repeated measurements yield slightly different values. Hence those working in this area, e.g. Budowle et al. (1991), consider their measurements to be quasi-continuous. In making certain calculations from databases of these values involving samples from specified populations, two genetic principles are invoked. The first is the Hardy-Weinberg equilibrium which amounts to statistical independence of the distribution of the pairs of alleles at a locus. The second is linkage equilibrium or statistical independence of the allelic distribution between different loci or more specifically mutual independence of the several loci used.

We bring to bear the methods devised to test these independence hypotheses within a locus and pairwise between loci. We note that while pairwise independence does not imply mutual independence, pairwise dependence implies mutual dependence.

The FBI has made data available that they use in forensic identification to illustrate these procedures. We used a Black database and several probes; D2S44 on 475 individuals, D1S7 on 359 individuals and

D17S79 on 550 individuals. Histograms of the fragment length data are given in Figures 1-3. In Table 1 the exchangeable non-identifiable quantile χ^2 tests are applied to each probe. At q = 2 we find the P-values to be .0346, .0134 and .0052 for D2S44, D17S79 and D1S7 respectively, which would indicate rejection of independence for all 3 loci. For testing independence between these three probes pairwise we first add the two fragment lengths within a probe for an individual. The database is such that there are only 342 individuals measured on both D2S44 and D1S7, 450 on both D2S44 and D17S79 and 336 on both D1S7 and D17S79 and we applied the test only to those pairs without missing values. In these cases the quantile tables for q = 2 are also given in Table 1. It is clear that this test would not provide evidence against independence. Of course the test may not have sufficient power to detect dependence if in fact the probes were dependent.

If the maternal and paternal alleles could be identified for each individual in the database, then we could apply the method for exchangeable and identifiable (X,Y). This would occur if the parents of each individual in the database were also profiled. However, this is rarely the case with databases such as the ones analysed here.

(Table 1 about here)

4. Remarks and Acknowledgement

These results can be extended for missing data situations as well.

The work here was supported in part by NIH Grant GM25271.

Appendix I

Under the null hypothesis X and Y are independent and (X_i,Y_i) is a random sample on (X,Y). We have shown that, conditional on the medians,

$$\Pr\left(N_{11}=n_{11}\right)=\frac{\binom{\frac{N}{2}}{n_{11}}\binom{\frac{N}{2}}{\frac{N}{2}-n_{11}}}{\binom{N}{\frac{N}{2}}} \qquad \text{(AI.1)}$$

$$E(N_{11})=\frac{N}{4} \text{ and } \mathrm{Var}(N_{11})\doteq\frac{N}{16}.$$

Now we will show that

$$\frac{4\left(N_{11}-\frac{N}{4}\right)}{\sqrt{N}}\rightarrow N(0,1).$$

Let d_N be an element of $\{4(i-N/4)/\sqrt{N}: i=0,1,\ldots,N/2\}$ such that $d_N \rightarrow d$ as $N\rightarrow\infty$. We will then show

$$\frac{\sqrt{N}}{4}\Pr\left[\frac{4\left(N_{11}-\frac{N}{4}\right)}{\sqrt{N}}=d_N\right]=\frac{e^{-\frac{d^2}{2}}}{\sqrt{2\pi}}.$$

This implies that

$$\frac{4\left(N_{11}-\frac{N}{2}\right)}{\sqrt{N}}\rightarrow N(0,1)$$

Gnedenko (1967, section 13). Using Stirling's approximation on AI.1 we obtain

$$\Pr\left[N_{11}=n_{11}\right] \doteq \frac{\left(\frac{N}{2}\right)^{2N+2}}{\sqrt{2\pi}\, n_{11}^{2n_{11}+1}\left(\frac{N}{2}-n_{11}\right)^{N-2n_{11}} N^{N+\frac{1}{2}}}.$$

Thus

$$\Pr\left[\frac{4\left(N_{11}-\frac{N}{4}\right)}{\sqrt{N}}=d_N\right]=\Pr\left[N_{11}=\frac{N}{4}+\frac{d_N\sqrt{N}}{4}\right]$$

$$\doteq \frac{\left(\frac{N}{2}\right)^{2N+2} N^{-\left(N+\frac{1}{2}\right)}}{\sqrt{2\pi}\left(\frac{N}{4}+\frac{d_N\sqrt{N}}{4}\right)^{\frac{N+d_N\sqrt{N}}{2}+1}\left(\frac{N}{4}-\frac{d_N\sqrt{N}}{4}\right)^{\frac{N-d_N\sqrt{N}}{2}+1}}$$

$$\doteq \frac{4}{\sqrt{2\pi N}}\, e^{-\frac{d^2}{2}}$$

as required. Further for q = 2,

$$X=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(N_{ij}-\frac{N}{4}\right)^2}{\frac{N}{4}}=\left(\frac{N_{11}-\frac{N}{4}}{\frac{\sqrt{N}}{4}}\right)^2$$

thus

$$X \to \chi_1^2.$$

Appendix II

Now for q = 2

$$X = \frac{8}{N}\left(N_{11} - \frac{N}{4}\right)^2 + \frac{4}{N}\left(N_{12} - \frac{N}{4}\right)^2 + \frac{4}{N}\left(N - 2N_{11} - N_{12} - \frac{N}{4}\right)^2$$

since $N_{21} = N - 2N_{11} - N_{12}$. However, the last term above can be rewritten as

$$\frac{24}{N}\left(N_{11} - \frac{N}{4}\right)^2 + \frac{8}{N}\left(N_{12} - \frac{N}{4}\right)^2 + \frac{16}{N}\left(N_{11} - \frac{N}{4}\right)\left(N_{12} - \frac{N}{4}\right).$$

Hence after some algebra, we obtain the quadratic form

$$X = \frac{16}{N}\left(N_{11} - \frac{N}{4}, N_{12} - \frac{N}{4}\right)\begin{pmatrix} 1 & -1 \\ -1 & 3 \end{pmatrix}^{-1}\begin{pmatrix} N_{11} - \frac{N}{4} \\ N_{12} - \frac{N}{4} \end{pmatrix}.$$

Hence we need only show that

$$\frac{4}{\sqrt{N}}\left(N_{11} - \frac{N}{4}, N_{12} - \frac{N}{4}\right) \to N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right) \qquad \text{AII.1}$$

where

$$\Sigma = \begin{pmatrix} 1 & -1 \\ -1 & 3 \end{pmatrix}$$

then $X \rightarrow \chi^2_2$ as N grows. We now proceed to demonstrate AII.1.

In the case where all (X_i,Y_i) i=1,...,N are exchangeable, identifiable and independent then we can show that

$$\Pr\left[N_{11} = n_{11}, N_{12} = n_{12}\right] = \frac{\binom{N}{n_{11},n_{12},n_{21},n_{22}}}{\binom{2N}{N}}$$

with the following constraints

$$n_{11} = n_{22} \quad \text{and} \quad n_{11} + n_{12} + n_{21} + n_{22} = N$$

leaving, say, N_{11} and N_{12} to vary.

Now we need to show that

$$\lim_{N\to\infty} \frac{N}{16} \Pr\left(\frac{N_{11} - \frac{N}{4}}{\frac{\sqrt{N}}{4}} = c_N, \frac{N_{12} - \frac{N}{4}}{\frac{\sqrt{N}}{4}} = d_N \right)$$

$$= \frac{|\Sigma|^{-\frac{1}{2}}}{2\pi} e^{-\frac{1}{2}(c,d)\Sigma^{-1}\binom{c}{d}}$$

where

$$c_N \in \left\{ \frac{i - \frac{N}{4}}{\frac{\sqrt{N}}{4}} : i = 0,\ldots,\frac{N}{2} \right\}$$

$$d_N \in \left\{ \frac{i - \frac{N}{4}}{\frac{\sqrt{N}}{4}} : i = 0,\ldots,N \right\}$$

and $c_N \to c$ and $d_N - d$ as N grows. This result is established by using the Stirling approximation for the factorials and some algebra in a fashion similar to Appendix I.

For the case where (X,Y) are not identifiable we have in addition to the above that $n_{12} + n_{21} = n_{12}^*$. Hence only N_{11} is free to vary and the result for q = 2 is just a special case of the above with

$$\frac{4\left(N_{11} - \frac{N}{4}\right)}{\sqrt{N}} \to N(0,1).$$

This result was also shown directly in Geisser and Johnson (1992).

References

Budowle, B. et al. (1991). Fixed bin analysis for statistical evaluation of continuous distribution of allelic data from VNTR loci for use in forensic comparisons. *American Journal of Human Genetics* **48** 841-855.

Geisser, S. and Johnson, W. (1992). Testing Hardy-Weinberg equilibrium on allelic data from VNTR loci. *American Journal of Human Genetics* **51** 1084-1088.

Gnedenko, B.V. (1967). *The Theory of Probability*. Chelsea, New York.

Figure 1. Histogram of 950 Fragment Lengths for D2S44

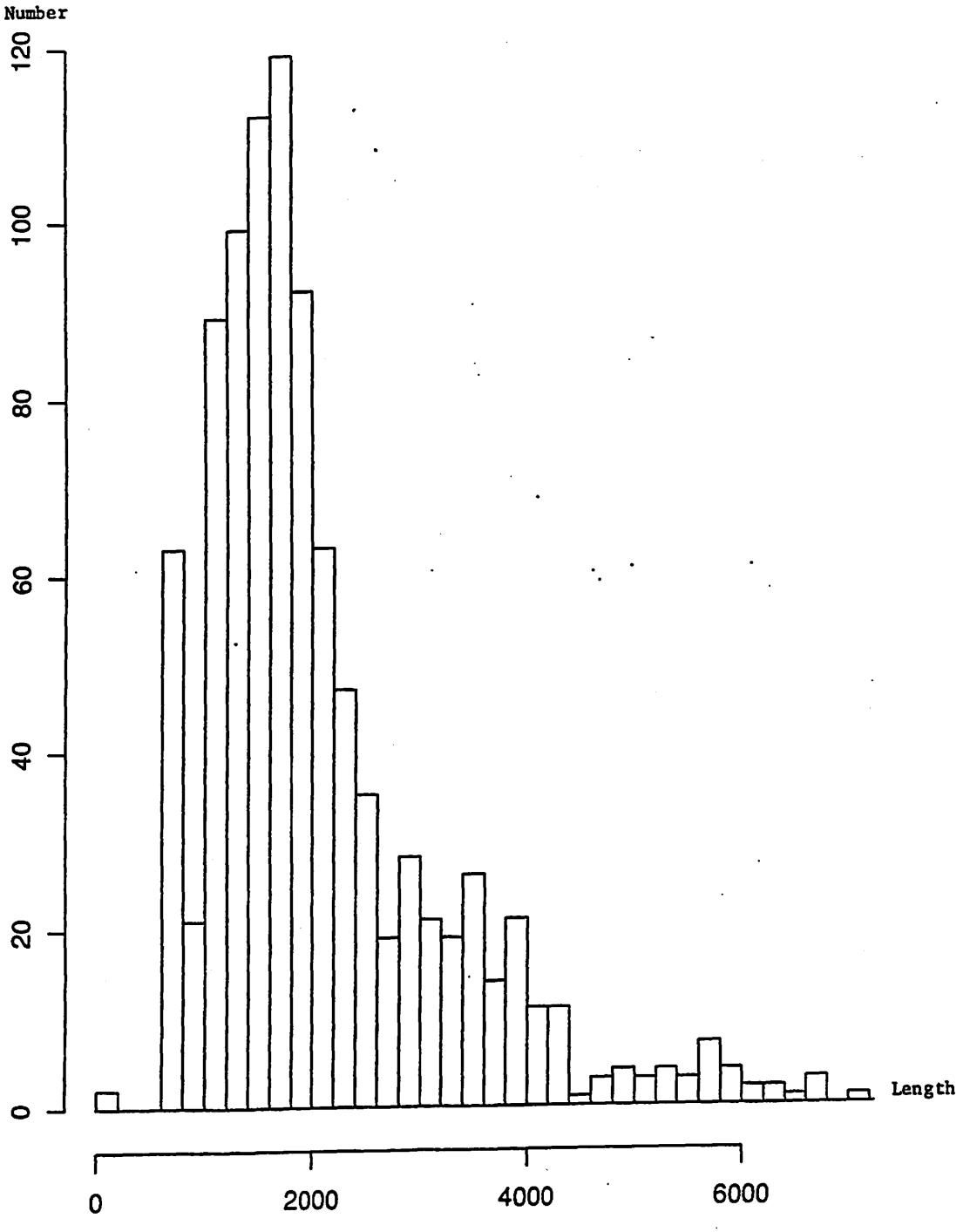

Figure 2. Histogram of 1100 Fragment Lengths for D17S79

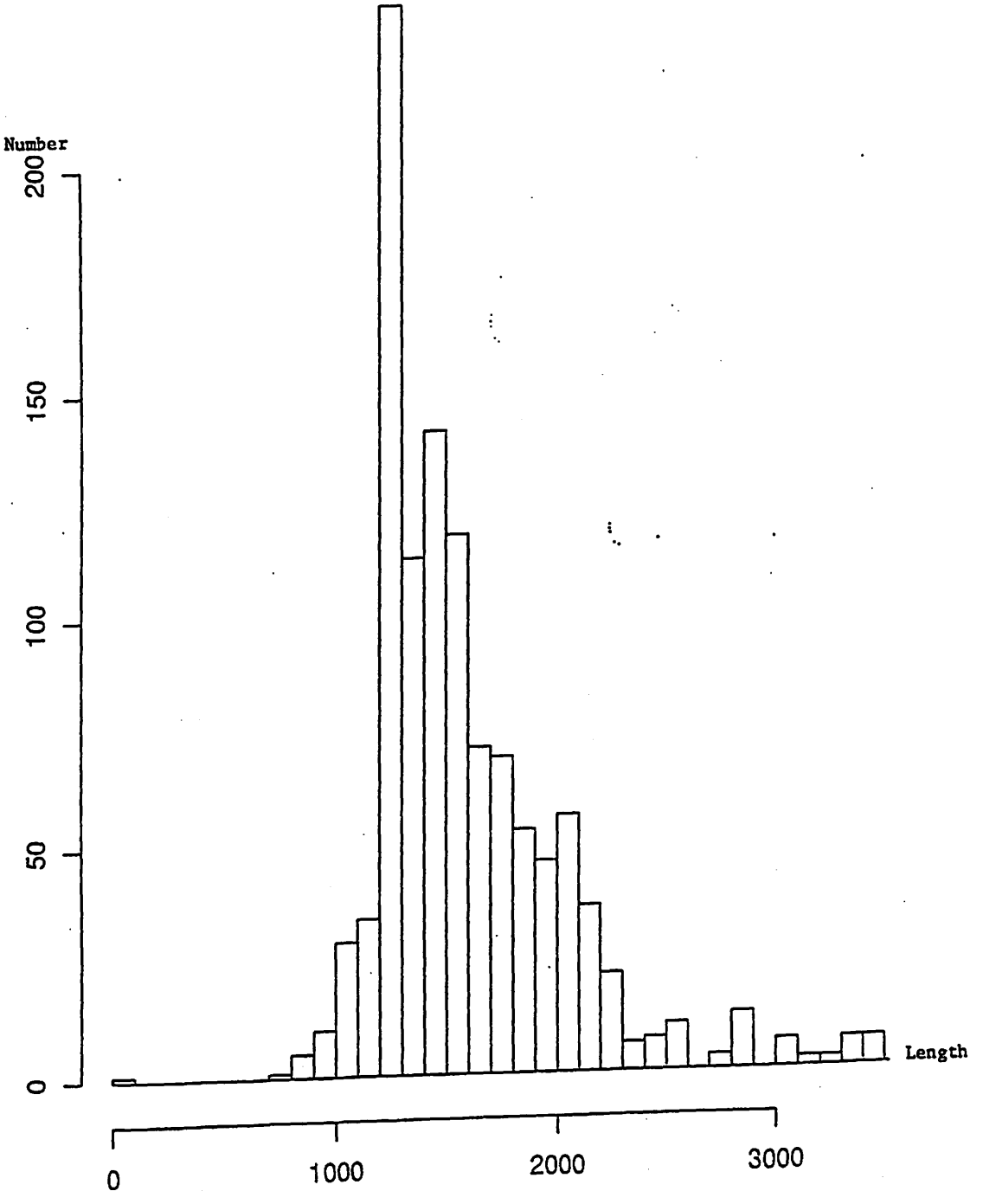

Figure 3. Histogram for 718 Fragment Lengths for D1S7

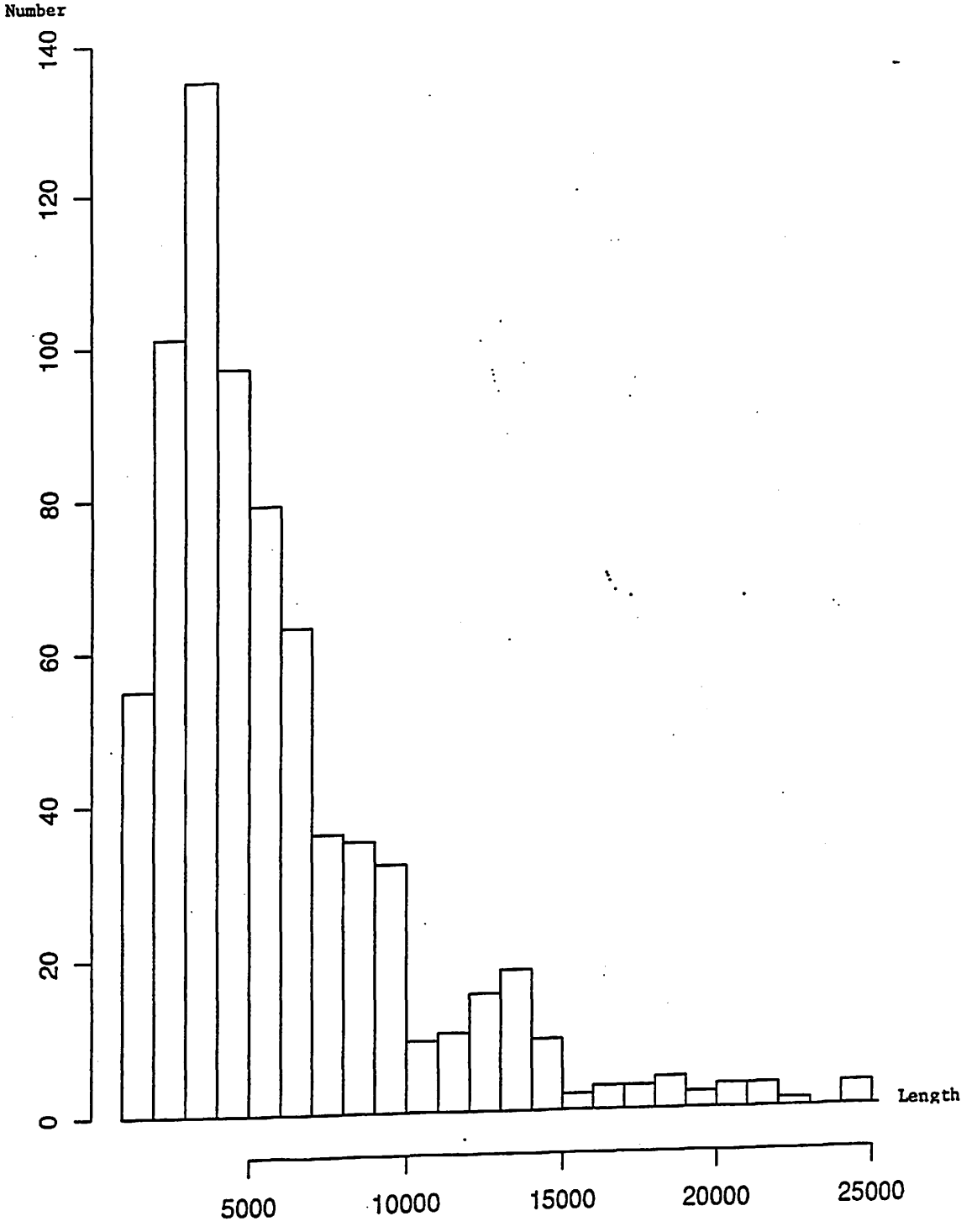

Table 1. Analysis of FBI Black Database Within and Between Probes D2S44, D17S79, D1S7 (q = 2); M = joint sample median, M_X, M_Y are individual sample medians; cells contain frequency and (expected frequency)

		< M	> M
D2S44	< M	130 (118.5)	214 (237)
N = 474			
X = 4.4641	> M		130 (118.5)
P = .0346			

		< M	> M
D17S79	< M	152 (137.5)	214 (237)
N = 550			
X = 6.1164	> M		152 (137.5)
P = .0134			

		< M	> M
D1S7	< M	103 (89.75)	153 (179.5)
N = 359			
X = 7.8245	> M		103 (89.75)
P = .0052			

		$< M_Y$	$> M_Y$
D2S44 × D17S79	$< M_X$	110 (112.5)	115 (112.5)
N = 450			
	$> M_X$	115 (112.5)	110 (112.5)

		$< M_Y$	$> M_Y$
D2S44 × D1S7	$< M_X$	86 (85.5)	85 (85.5)
N = 342			
	$> M_X$	85 (85.5)	86

		$< M_Y$	$> M_Y$
D17S79 × D1S7	$< M_X$	82 (84)	86 (84)
N = 336			
	$> M_X$	86 (84)	82 (84)